Course materials

Your points

This course has already ended.

Submit

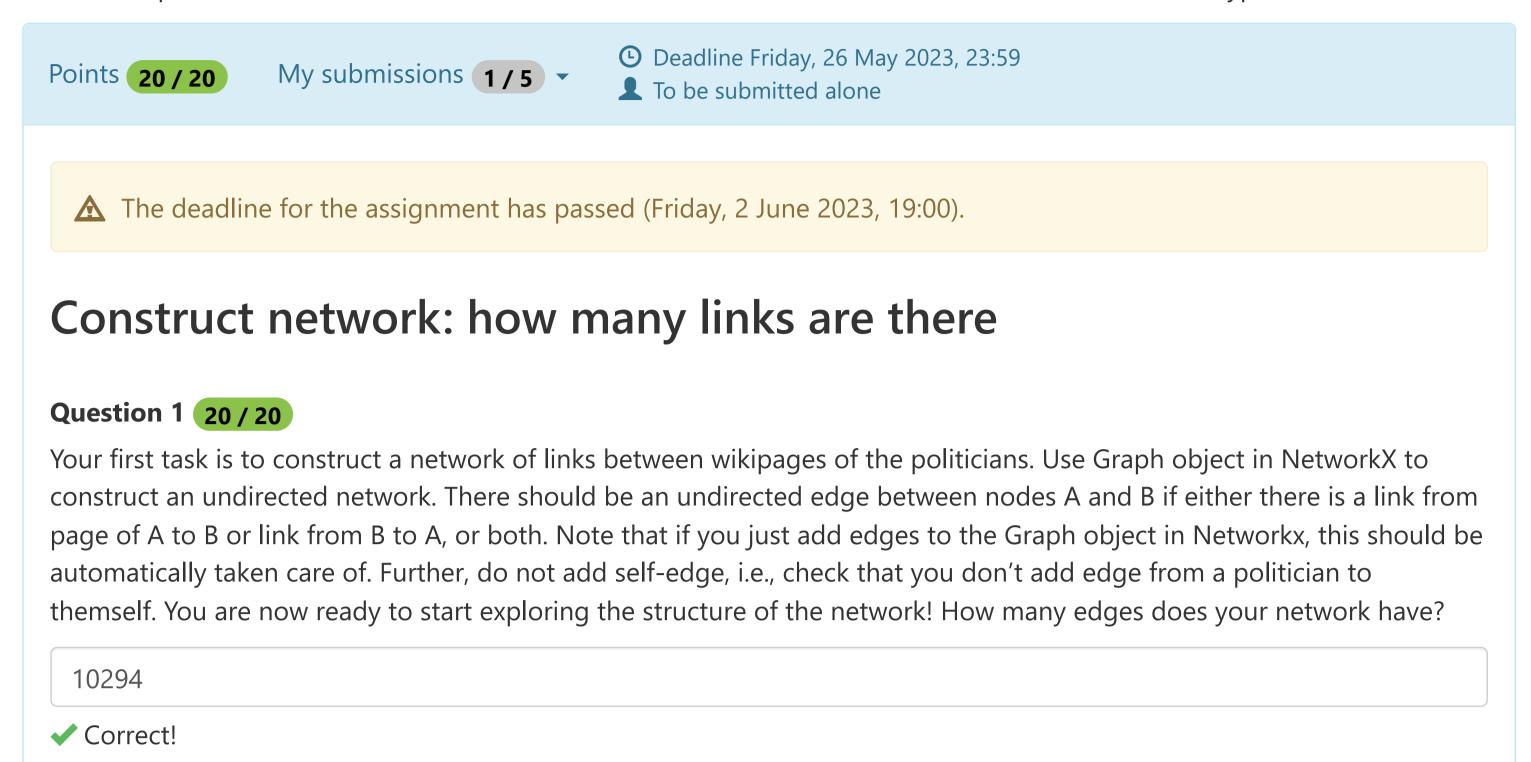
« 6.1 The project Course materials 6.3 Returning the project »

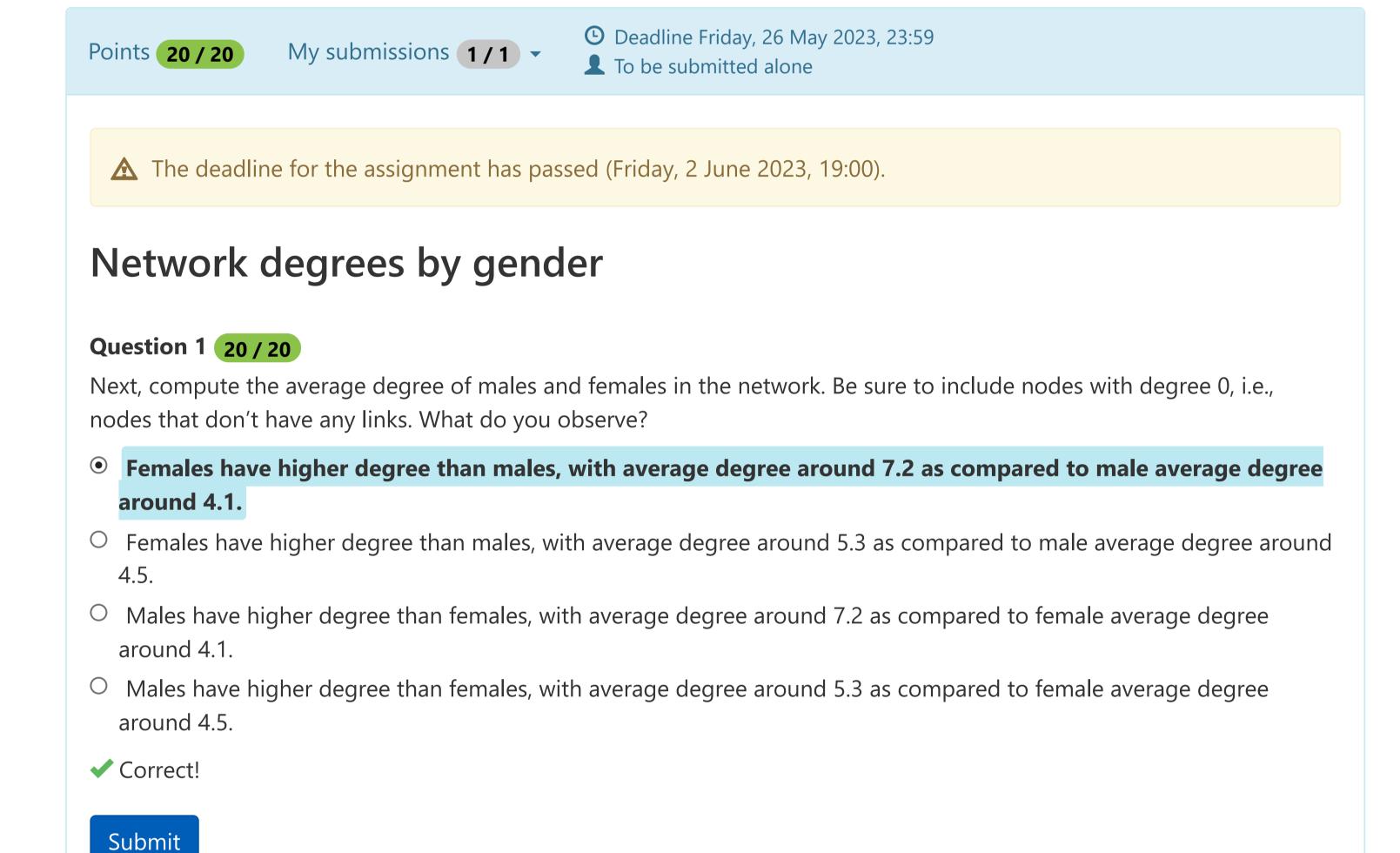
CS-E4730 / 6. Week: Introduction to the project / 6.2 Getting started with data

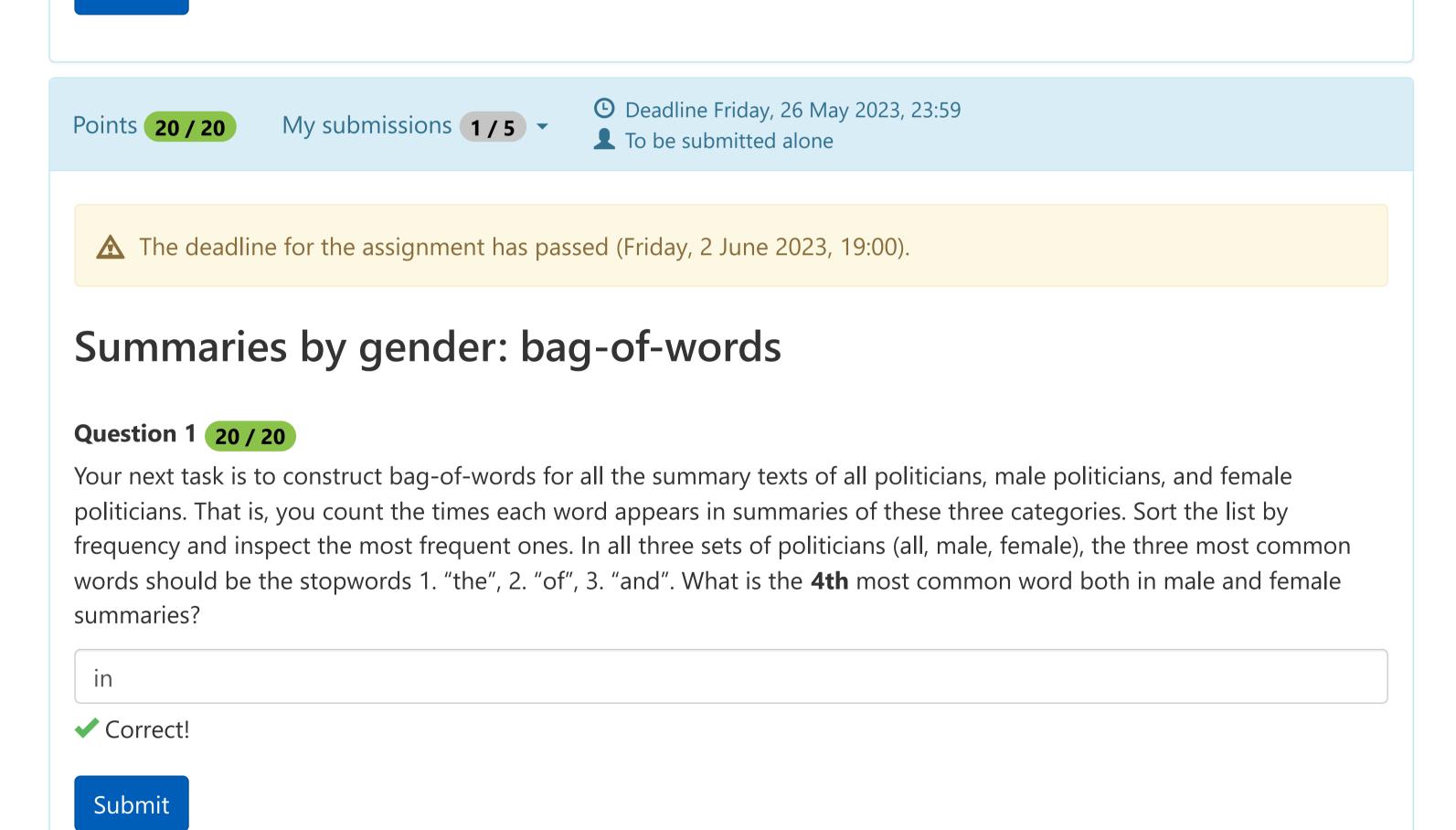
Getting started with data

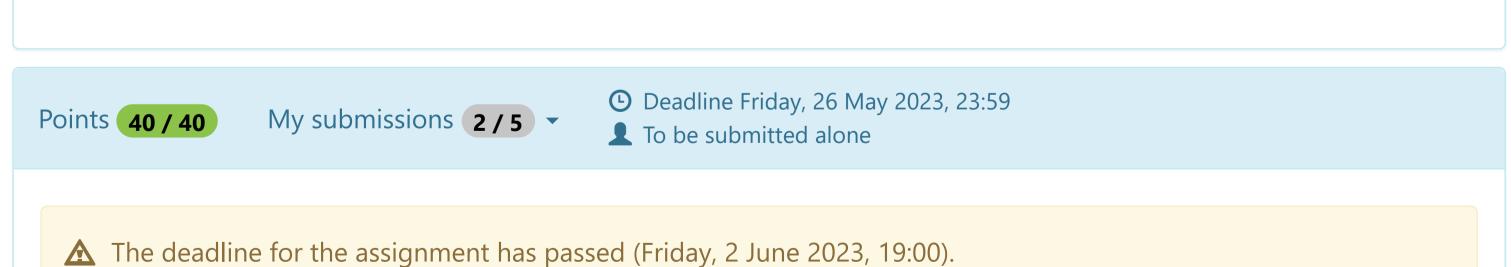
The purpose of this initial exercise is to get started with analysing the data. Download the notebook and the data file for politicians in Wikipedia, and answer the following questions. The idea of these questions is to check that everything is technically correct and you are on the right track. In the project you can try out more ideas and interpret the results.

Download the pre-fetched data here politicians.json and the template notebook here project-wikipedia.ipynb. Note that as this notebook is meant for the project, it is less structured than a typical notebook we have set up for exercises. You might need to do some exploration on how the data is formatted and write bit more code from the scratch than in a typical exercise.









Summaries by gender: tf-idf

Question 1 40 / 40

Your final task is to compute the average tf-idf scores for words in each category. To do this, first compute the average tf score by simply dividing the bag-of-words number you computed in the previous question by the total number of politicians of a certain kind. For example, if there are 386 female politicians, and the word "the" appears 2293 times in their summaries, the tf value for "the for females is 2293/386. The idf score you count by first computing the document frequency df, which is the number of summary texts wherein the word appears at least once and divide that by the total number of politicians. For example, if the word "the" appears in 6149 of all 6880 summary texts then the df value is 6149/6880. The inverse document frequency score, i.e., the idf score, is then the logarithm of the inverse of the df, for example for the word "the" idf = log(6880/6149). Use the natural logarithm here. The tf-idf score is thus the multiplication of these two scores, that is, tf*idf, and its value for females and the word "the" is 2293/386 * log(6880/6149). Inspect the top tf-idf scores for males and females. For females the top score is given by "she". What is the word with **second highest** score for females?

```
her

Correct!

Submit
```

« 6.1 The project Course materials 6.3 Returning the project »